

THE PROBABILITY DISTRIBUTION OF NEAREST NEIGHBOR POINT
CONFIGURATIONS OF A STATIONARY PROCESS ON THE REAL LINE

BU-211-M

D. S. Robson

February, 1966

Abstract

A given point a_n taken from the monotonically increasing sequence $\{a_i\}$, $-\infty < i < \infty$, will be said to generate a k -point nearest neighbor configuration if the path from a_n to its nearest neighbor $a_n^{(1)}$ ($a_n^{(1)}$ is either a_{n-1} or a_{n+1}), from $a_n^{(1)}$ to its nearest neighbor $a_n^{(2)}$, etc., includes exactly k points of the sequence $\{a_i\}$. If the differences $a_{i+1} - a_i$ are the realizations of independent and identically distributed (non-negative) continuous random variables then the cumulative probability distribution of k is

$$P(K \leq k) = 1 - \frac{2}{(k+1)!} \quad .$$

THE PROBABILITY DISTRIBUTION OF NEAREST NEIGHBOR POINT
CONFIGURATIONS OF A STATIONARY PROCESS ON THE REAL LINE

BU-211-M

D. S. Robson

February, 1966

Introduction

Standard methods of testing for "random dispersion" of an organism in the field consists of counting organisms per quadrat and testing the goodness-of-fit of the observed quadrat counts to a Poisson distribution. This rather tedious counting method is now being supplanted by a variety of "nearest neighbor methods", whereby a sample of organisms is randomly chosen and the distance is measured from each selected organism to its nearest neighbor. The Poisson model (in the plane) implies that these nearest-neighbor distances are independent and identically distributed as a scalar multiple of a χ^2_2 variable (χ^2_2 = chi-square on 2 degrees of freedom), and the model is tested accordingly. Extensions of this method include measuring also the distance to the second nearest neighbor, the third nearest neighbor, and so on.

Such methods also provide the data for point and interval estimates of the parameter λ of the underlying Poisson process; however, if the sole objective is to test for random dispersion then this sample information concerning the value of λ is irrelevant. The cost of obtaining this superfluous information is the added cost of measuring absolute distances as compared to relative distances between neighbors, since relative distances are independent of density.

Biometrics Unit, Plant Breeding Department, Cornell University.

One sampling procedure which eliminates this unnecessary cost consists of counting types of nearest neighbor configurations generated by the selected sample. A configuration type k is characterized by the number k of organisms included in a chain which links a selected organism A_1 to its nearest neighbor A_2 , links A_2 to its nearest neighbor A_3 , links A_3 to its nearest neighbor A_4 , and so on, terminating at the point A_k where (for the first time in the chain) the nearest neighbor to A_k is A_{k-1} . The empiric frequency distribution of the configuration types generated by n selected organisms may then be compared with the theoretical distribution (which does not depend upon λ) to provide a test of the Poisson model. The present note will develop this theoretical frequency distribution of configuration type for a Poisson process on the real line; in this case, however, the goodness-of-fit test will be seen to be powerless against stationary alternatives to the Poisson process. In other words, the goodness-of-fit test will be a test of the stationary property rather than the Poisson property.

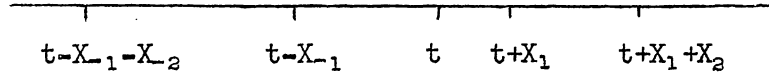
Nearest Neighbor Configurations on the Line

For convenience we shall regard the real line as a time axis, and let

$$F(x|t) = P \left(\begin{array}{c|c} \text{the next event will occur} & \text{the last event has occurred} \\ \text{before time } t+x & \text{at time } t \end{array} \right)$$

For a Poisson process, $F(x|t) = 1 - e^{-\lambda x}$, where λ is the expected number of events occurring during a unit time interval. The Poisson process is stationary in the sense that $F(x|t)$ does not depend on t . All processes for which the waiting times between successive events are independent and identically distributed (non-negative) random variables will generate the same frequency

distribution of configuration types. For if t is the time when a randomly selected event occurred then the probability that this selected event generates a two-point configuration:



is the probability that the distance from the point t to the nearest neighboring point (distance to nearest neighbor = $\min(X_{-1}, X_1)$) is less than the distance from that nearest neighbor to any other point. Thus,

$$\begin{aligned}
 p_2 &= P\{\min(X_{-1}, X_1) < X_2\} = P\{\min(X_{-1}, X_1) < X_2\} \\
 &= 1 - P\{X_2 < \min(X_{-1}, X_1)\} \\
 &= 1 - P\{X_2 = \min(X_{-1}, X_1)\} \\
 &= 1 - \frac{1}{3}
 \end{aligned}$$

Similarly, the probability of a three-point configuration is

$$\begin{aligned}
 p_3 &= P\{X_2 < \min(X_{-1}, X_1), X_3 > X_2\} \\
 &= P\{X_2 < \min(X_{-1}, X_1)\} [1 - P\{X_3 < X_2 \mid X_2 < \min(X_{-1}, X_1)\}] \\
 &= P\{X_2 < \min(X_{-1}, X_1)\} [1 - P\{X_3 < \min(X_{-1}, X_1, X_2)\}] \\
 &= \frac{1}{3} (1 - \frac{1}{4})
 \end{aligned}$$

and, in general, the probability of a k -point configuration is

$$\begin{aligned}
 p_k &= P\{X_{k-1} < X_{k-2} < \dots < X_2 < \min(X_{-1}, X_1), X_k > X_{k-1}\} \\
 &= P\{X_2 < \min(X_{-1}, X_1)\} P\{X_3 < \min(X_{-1}, X_1, X_2)\} \dots \\
 &\quad \dots P\{X_{k-1} < \min(X_{-1}, X_1, X_{k-2})\} [1 - P\{X_k < \min(X_{-1}, X_1, \dots, X_{k-1})\}] \\
 &= \frac{1}{3} \left(\frac{1}{4}\right) \dots \left(\frac{1}{k}\right) \left[1 - \frac{1}{k+1}\right] \\
 &= \frac{2k}{(k+1)!}
 \end{aligned}$$

The cumulative form of this distribution is

$$P_k = p_2 + p_3 + \dots + p_k = 1 - \frac{2}{(k+1)!}$$

as may be readily confirmed by finite induction.